Robust Estimation of Biometric Data

Guido Kuijper

Abstract: This study presents the design of a robust estimator $\hat{C}$ to apply on biometric data analysis involving facial recognition. This functional estimates the covariance matrix of a multivariate Gaussian distribution by seperately estimating the matrix elements. It is first mathematically derived, then classified by means of its efficiency at Gaussian distributions and finally applied to both synthetic and real biometric data. The synthetic experiments show the $\hat{C}$-estimator performs in between the sample covariance and the MCD estimator. The test with the real data shows clear improvement of the robust $\hat{C}$-estimator as it was able to link two faces for which the sample covariance estimator was not able to.

1 Introduction

Nowadays one of the fields that employ statistical methods the most is the field of biometrics. This application is referred to as biostatistics and within biometrics, biostatistics is often applied to genetical data to model the influence of groups of genes. [1]
Another application of biostatistics is in the field of computer vision and in particular the field of facial recognition. This paper complements the method presented in Spreeuwers [2] and improved in Spreeuwers [3]. These papers use a classifier based distinction between the classes and class members, where it is assumed both the classes and members of a class are normal distributed.

The covariance matrices of these multivariate distributions are estimated, which are then used in a LDA-based classification. The estimation of covariance matrices is notoriously susceptible to outlier interference [4]. To solve this problem, one can turn to *robust* statistics. This branch deals with outliers and corrupted data. Robust estimation has seen a good amount of mathematical development starting from Hampel [5], who defined the concept of robustness from the influence function, called the influence curve back then, and some measures that can be derived from it. A public start on the robust estimation of covariance matrices was made by Gnanadesiken and Kettenring [6]. Maronna [7] was the first to formulate robust M-estimators in this context. Subsequently many approaches to this problem were described, including the *minimum determinant covariance* (MCD) [8], projection methods [9] and computing the smallest volume ellipsoid over the set of data [10].

Overview : This paper starts with a short overview of important definitions in robust statistics in section 2, the description of the often-used MCD estimator in section 3 and in section 4 the design of the $\hat{C}$-estimator is presented. Then in section 5 experiments are done with synthetic data that model often occuring contaminations in facial datasets. A score evaluation of real facial data can be found in section 6. The paper ends with a discussion of the results in section 7.

2 Robust Statistics

This section was written as a summary of Hampel et al [11], with the intention of introducing some important concepts that are going to be discussed in this paper. One of the most important definitions in robust statistics is the definition of the *influence function* (IF).

Definition : The influence function of functional T at distribution F is given by:

$$IF(x;T,F) = \lim_{t\to 0} \frac{T((1-t)F + t\Delta_x) - T(F)}{t}$$

The influence function is therefore a directional directive from T at F to T at Δ_x with the latter denoting the probability measure that puts mass 1 at the point x. Heuristically, the interpretation of the influence function is that it is a quantity of how much the estimator T becomes corrupted when sample x is contamined with Δ_x.
The influence function is chiefly a heuristic tool. It also is at the basis of a few other statistical tools and quantities, such as the *gross-error sensitivity*:

Definition : The gross-error sensitivity γ of a functional T at distribution F is given by:

$$\gamma(T,F) = \sup_x |IF(x;T,F)|$$

If this quantity is finite, we call the estimator B-robust at F. Intuitively, this means no contamination at any sample results in an unbounded error in the estimation. It is a desirable property for an estimator to be B-robust.

Theorem : The asymptotic variance of T at F equals:

$$V(T,F) = \int IF(x;T,F)^2 dF(x)$$

This gives a way to quantify the variance of the estimator in the asymptotic domain, which relies on the notion of the influence function. The asymptotic variance is bounded by the inverse of the Fisher information. The Fisher information is defined as:

Definition : The Fisher information at F_θ equals:

$$I_F(f_{X,\theta}) = \int f_{X,\theta}(x) \left(\frac{\partial \ln f_{X,\theta}(x)}{\partial \theta} \right)^2 dx$$

It provides a measure of how much information a random variable carries of a parameter θ that its distribution depends on.
The fundamental bound of the asymptotic variance is given by the Cramer-Rao inequality:

$$V(T_\theta, F) \geq \frac{1}{I_F(f_{X,\theta})}$$

Using this bound, we can define the asymptotic efficiency of an estimator:

Definition : The asymptotic efficiency of an estimator T of a parameter θ at distribution F is given by: Eff = $(\mathrm{V(T,F)} I_F(f_{X,\theta}))^{-1}$.

Lastly we have an independent quantity named the maximum breakdown point (p). This is the maximum amount of observations before the estimation can go arbitrarily wrong. Intuitively it makes sense that p can never be greater than 0.5, as in that case the outliers would be considered the data.

3 MCD Estimator

Definition: Consider a set of data $X = [x_1, ..., x_n]$ with each x_i consisting of p different observations. Therefore X is a n x p matrix. Using this dataset we can derived the *Malahanobis distance* [12] which is defined to be the set of p-dimensional points which satisfy:

$$MD(x) = \sqrt{(x-\bar{x})^t S^{-1} (x-\bar{x})} = \sqrt{\chi^2_{p,0,975}}$$

Where $\bar{x}$ is the sample mean and S the sample covariance. With χ^2_α the α-quantile of the chi-squared distribution is meant.
The MCD has a parameter h. For the h observations that minimize the determinant of the sample covariance matrix S, we can derive two estimates:

1. $\hat{\mu}_0$ is the sample mean of these h observations.
2. $\hat{\Sigma}_0$ is the corresponding covariance matrix (multiplied with a scalar to be consistent).

These estimates yield another distance metric:

$$d_i = \sqrt{(x-\hat{\mu}_0)^t \hat{\Sigma}_0^{-1} (x-\hat{\mu}_0)}$$

This distance d_i is then used to find the MCD estimates for location and scatter:

$$\hat{\mu}_{\mathrm{MCD}} = \frac{\sum_{i=1}^n W(d_i^2) x_i}{\sum_{i=1}^n W(d_i^2)}$$

$$\hat{\Sigma}_{\mathrm{MCD}} = c \frac{1}{n} \sum_{i=1}^n W(d_i^2)(x_i - \hat{\mu}_{\mathrm{MCD}})(x_i - \hat{\mu}_{\mathrm{MCD}})^t$$

The constant c is for consistency.
W is a weighting function and is defaulted to $W(d^2) = I\left(d^2 \leq \sqrt{\chi^2_{p,0,975}}\right)$ in MATLAB [8].

Properties : The MCD estimator was shown to be B-robust [13]. In addition, the maximal breakdown value is (n-p+2)/2 $\forall h, (n+p)/2 \leq h \leq (n+p+1)/2$. A common issue with robust estimators is the trade-off between asymptotic variance/efficiency and robustness [11]. In the case of the MCD estimator this issue manifests in the choice for h. By denoting the mass of the data not determining the MCD by $0 < \beta < 1$, Croux and Haesbroeck [13] demonstrated that the

efficiency for Gaussian distribution decreases if β increased. However, for increasing amounts of observations p, the efficiency rises. They proved for the MCD estimator:

$$\lim_{p\to\infty} \text{Eff}(\hat{\Sigma}_{\text{MCD},ij}, \Phi) = 1 - \beta \quad \forall i \geq 1, j \leq p$$

In addition, it was shown that a class of estimators, which include $\hat{\Sigma}_{\text{MCD}}$, have asymptotically the same properties as [14]. Lastly, the FAST-MCD algorithm was developped by Rosseeuw and Van Driessen [15]. This theoretical clarity and application support of the MCD estimator make it an attractive choice.

4 $\hat{C}$-Estimator

A second choice for a robust estimator of covariance is the $\hat{C}$-estimator. This estimator was proposed by Gnanadesikan and Kettenring [6]. They propose to estimate each seperate element in the covariance matrix based on the following identity, valid for random variables X and Y:

$$\text{Cov}(X,Y) = \frac{1}{4}\left(\text{var}(X+Y) - \text{var}(X-Y)\right)$$

By plugging in a robust estimator of scale, denoted by $S(X)$, we can estimate each covariance by:

$$\hat{C}(X,Y) = \frac{1}{4}\left(S(X+Y)^2 - S(X-Y)^2)\right)$$

This estimator has some problems:

- A covariance matrix is necessarily positive-semidefinite [16]. However this estimator does not automatically produce a positive-semidefinite matrix. Therefore the shape ellipsoid may be hyperboloid.
- As far as the author knows, no robustness properties have been derived for this estimator.

The rest of this section is devoted to solving these problems:

Conditioning the $\hat{C}$-estimator : A covariance matrix is only positive semi-definite if one of the random variables is linearly dependent on another. Therefore, for data where one does not expect linear dependence between samples, we will introduce a scheme to enforce a positive definite (PD) covariance matrix. We use the following theorem, by Sylvester [17]:

Theorem 1. *: A matrix $M \in \mathbb{R}^{nxn}$ is PD if and only if all of its principle minors are PD and det(M) > 0.*

This theorem is also called Sylvester's criterion. This criterion is used to build the matrix in a way that resembles mathematical induction:

Theorem 2. *: A matrix $M \in \mathbb{R}^{1x1}$ is PD if and only if det(M) > 0.*

The covariance matrix of a single variable X is given by the variance of X or cov(X,X). As cov(X,X) > 0, a covariance matrix of one variable is always PD.

Now suppose we have n variables and the covariance matrix of p variables with p < n is given by $A_p \in \mathbb{R}^{pxp}$ and $\det(A_p) > 0$. We have then that the covariance matrix of p+1 variables has the following form:

$$\hat{\Sigma}_p = \begin{bmatrix} A & B \\ B^T & D \end{bmatrix}$$

With $\hat{\Sigma}_p$ the sample covariance matrix of p+1 variables. B is given by $\hat{C}(X_i, X_{p+1})\ \forall i < p+1$ and B is given by $\hat{C}(X_{p+1}, X_{p+1})$, T denoting the transpose. Notice that if $\det(\hat{\Sigma}_p) > 0$, the estimated covariance matrix will be PD.
For the determinant of a matrix M, the following identity is the case:

$$\det\begin{bmatrix} A & B \\ C & D \end{bmatrix} = (D+1)\cdot\det(A) - \det(A+BC)$$

For $C = B^T$ we have an expression for $\det(\hat{\Sigma}_p)$. Another determinant identity, also by Sylvester [18], is as follows:

$$\det(A+BC) = \det(A)\cdot\det(I_m + C\cdot A^{-1}\cdot B)$$

For $C = B^T$ we substitute this identity in 4, obtaining:

$$\begin{aligned}\det(\hat{\Sigma}_p) &= \det\begin{bmatrix} A & B \\ B^T & D \end{bmatrix} \\ &= \det(A)\left((D+1) - \det(I_m + B^T A^{-1} B)\right)\end{aligned}$$

We require $\det(\hat{\Sigma}_p) > 0$. Therefore the following inequality needs to hold:

$$\det(A)\,(D+1) > \det(A)\det(I_m + B^T A^{-1} B)$$

As $\det(A) \neq 0$, we can divide over it to obtain.

$$(D+1) > \det(I_m + B^T A^{-1} B)$$

As B is a row vector and B^T a column vector, we are taking the determinant over a number, which is the number itself:

$$(D+1) > 1 + B^T A^{-1} B \rightarrow D > B^T A^{-1} B$$

Therefore if we multiple D with a scalar c, we can enforce our estimated matrix to be PD. We lose unbiasedness for the diagonal elements of the however. By minimizing c, this problem can be partially circumvented.

Robustness properties of the $\hat{C}$-estimator To establish robustness properties, as suitable candidate for the robust scale estimator S(x) has to be defined first. An attractive candidate is the Q_n estimator, as described by Rosseeuw and Croux [19].

$$Q_n = d\left(|x_i - x_j|; i < j\right)_{(k)}$$

Where x_i, x_j are samples , (k) denotes the k-th order statistic and d is a consistency factor equal to 2.2219 at Gaussian distributions.

This estimator was shown to have a breakdown point of 50% of the samples. Additionally, it has an efficiency at Gaussian distributions of 82.27%. The functional of this estimator at a distribution F is given by:

$$Q(F) = \inf\left(s > 0; \int F(t + d^{-1}s)dF(t) \geq 5/8\right)$$

The influence function of the estimator at distribution F is given by:

$$IF(x; Q, F) = d\frac{1/4 - F(x + d^{-1}) + F(x - d^{-1})}{\int f(y + d^{-1})f(y)dy}$$

With f the density of F.
This influence function describes a smooth curve. Now the $\hat{C}$-estimator can be defined as follows:

$$\hat{C}(X,Y) = \frac{1}{4}\left(Q_n\left(X+Y\right)^2 - Q_n\left(X-Y\right)^2\right)$$

To establish robustness properties of the $\hat{C}$-estimator, we will need to establish its influence function. The estimator is dependent on two different variables, which means that we have to determine the *partial* influence functions instead, denoted by PIF(x; T, F). These partial influence functions were introduced by Hampel et al [11] in the form of influence functions for tests. Partial influence functions in their current form were defined by Pires and Branco [20] However, the partial influence functions PIF(x;Q_n,F) and PIF(y;Q_n,F) are hard to directly evaluate, as Q_n is not linear. Therefore we will perform a change of variables: let Z = X + Y and W = X - Y. The $\hat{C}$-estimator now reads as follows:

$$\hat{C}(X,Y) = \frac{1}{4}\left(Q_n(Z(X,Y))^2 - Q_n(W(X,Y))^2\right)$$

Notice Z and W are normally distributed.
If $X, Y \sim N(0,1)$, $Z \sim N(0, \sigma_+)$ and $Z \sim N(0, \sigma_-)$.

Here $\sigma_\pm = \sqrt{2 \pm 2 \cdot \mathrm{cov}(X,Y)}$. As the influence function is special case of a Gâteaux derivative [21], we are allowed to use the chain rule if the influence function is continuous. Therefore:

$$PIF(z; \hat{C}, F_z) = \frac{1}{2} \cdot Q(F_z) \cdot IF(z; Q_n, F_z)$$
$$PIF(w; \hat{C}, F_w) = -\frac{1}{2} \cdot Q(F_w) \cdot IF(w; Q_n, F_w)$$

Where Q(F) denotes the estimator functional acting on a distribution F.
Both partial influence functions are B-robust, as the gross-error sensitivity is finite.
Theorem 3. *: $\gamma(\hat{C}, \Phi) = \sup_x |PIF(x; \hat{C}, \Phi)|$ and $\gamma(\hat{C}, \Phi) = \sup_y |PIF(y; \hat{C}, \Phi)|$ are finite.*

Proof. : As z = x+y is normally distributed, $\frac{z}{\sigma_+}$ is distributed N(0,1), therefore $PIF(\frac{x+y}{\sigma_+}; \hat{C}, \Phi) = PIF(z; \hat{C}, F_z)$. As the latter is a bounded function, set x = 0 or y = 0. □

Intuitively this is correct, since $PIF(z; \hat{C}, F_z)$ consists of possible point contaminations in x or y, but any point contamination will be the same as a point contamination in z.

Asymptotic Properties : As we aim to evaluate the $\hat{C}$-estimator for the standard normal distribution Φ, we need a change of variables. In a way analogous to the proof of 3, it can be seen that $\frac{z}{\sigma_+}$ and $\frac{w}{\sigma_-}$ are both distributed N(0,1).

Using this change of variables, we obtain:

$$PIF\left(\frac{z}{\sigma_+};\hat{C},\Phi\right) = \frac{1}{2}\cdot Q(\Phi)\cdot IF\left(\frac{z}{\sigma_+};Q_n,\Phi\right)$$
$$PIF\left(\frac{w}{\sigma_-};\hat{C},\Phi\right) = -\frac{1}{2}\cdot Q(\Phi)\cdot IF\left(\frac{w}{\sigma_-};Q_n,\Phi\right)$$

Note that $Q(\Phi) = 1$ so it is no longer considered in the derivation.
We use the definition of asymptotic variance as given by Pires and Branco [20] which is to be evaluated for the $\hat{C}$-estimator at a standard normal distribution for X and Y, with the additional case that we have as many samples for X as for Y:

$$V(\mathrm{c};\hat{C},\Phi_1,\Phi_2) = 2\cdot V_1(\mathrm{c};\hat{C},\Phi_1,\Phi_2) + 2\cdot V_2(\mathrm{cov};\hat{C},\Phi_1,\Phi_2)$$

Where the partial asymptotic variances V_i are given by:

$$V_i(\mathrm{c};\hat{C},\Phi_1,\Phi_2) = \int IF_i(x;\hat{C},\Phi_1,\Phi_2)^2 d\Phi_i(x)$$

We end up with the following expression for the asymptotic variance:

$$V(\mathrm{c};\hat{C},\Phi_1,\Phi_2) = \frac{1}{2}\int IF\left(\frac{z}{\sigma_+};\hat{C},\Phi\right)^2 d\Phi\left(\frac{z}{\sigma_+}\right) + \frac{1}{2}\int IF\left(\frac{w}{\sigma_-};\hat{C},\Phi\right)^2 d\Phi\left(\frac{w}{\sigma_-}\right)$$

V can now be evaluated with help of numerical integration. Notice that V in fact depends on the covariance that is being estimated. Therefore we now have to plot V versus the covariance. This plot can be found in figure 1. By a probatilistic corollary of the Cauchy-Swartz inequality, we have:

$$|\mathrm{Cov}(X,Y)|^2 \leq \mathrm{Var}(X)\cdot\mathrm{Var}(Y)$$

Since we are seeking the variance at Φ, the variances of X and Y are 1 and $-1 \leq \mathrm{Cov}(X,Y) \leq 1$.
The asymptotic variance appears to be almost constant over the whole covariance spectrum. The value is about 0.6089, which is very close to the asymptotic variance of Q_n as reported by Croux and Rosseeuw [19].

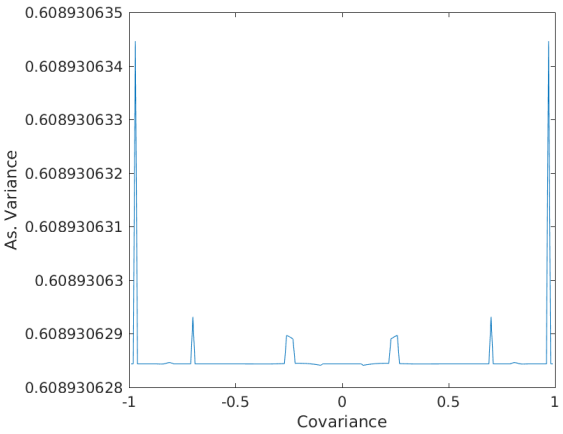

Figure 1: Asymptotic Variance versus Covariance

To evaluate the efficiency at Gaussian distributions, a different kind of Fisher information is required: the joint Fisher information [22]. This quantity is dependent on the joint density function $f_{X,Y}(x,y)$ and the parameter Θ. It is defined to be:

$$I_F(f_{X,Y,\Theta}) = \iint_{X,Y,\Theta}(x,y)\left(\frac{\partial \ln f_{X,Y,\Theta}(x,y)}{\partial\Theta}\right)^2 dxdy$$

In our case, the parameter Θ is given by Cov(X,Y). Note that Cov(X,Y) = Cov(Y,X). Since the Fisher information gives a score of the amount of information a random value carries about an unknown parameter Θ, an observable value in this case carries twice as much information since it has a score on Cov(X,Y) and on Cov(Y,X). The efficiency at Gaussian distributions is then given as follows:

$$\mathrm{Eff}(c;\hat{C},\Phi,\Phi) = \frac{1}{\left(2I_F(f_{X,Y,\mathrm{Cov}(X,Y)})V(\mathrm{c};\hat{C},\Phi_1,\Phi_2)\right)}$$

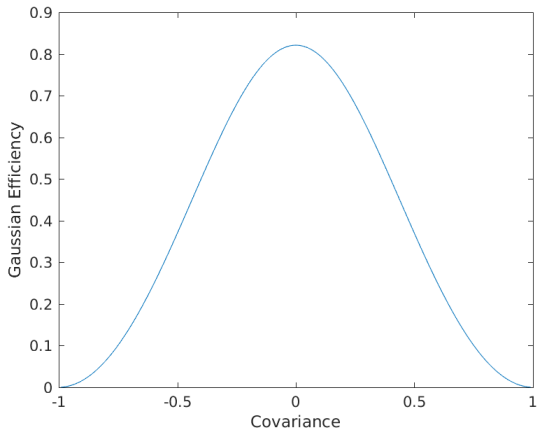

Figure 2: Gaussian Efficiency versus Covariance

This can be numerically solved as well and figure 2 is obtained. We see that the $\hat{C}$-estimator has a high efficiency for low covariance (82.11% at the peak) but quickly loses efficiency when estimating higher covariances. 82.11% is very close to the Gaussian efficiency of the Q_n estimator, which sits at 82.27%. By removing the strange peaks from figure 1, we get a peak efficiency of 82.28%.

5 Sample Asymptotic Variance

In the context of facial recognition, we would like to find out the performance of the MCD- and $\hat{C}$-estimators in comparison to the normal sample covariance matrix S. To obtain this knowledge, various experiments with synthetic data have been performed. In lieu of the synthetic data portraying as actual facial data, the PD-enforcing scheme has not been utilised. This has four reasons:

1. Facial data often contains pixels that are linearly dependent on their neighbour pixel, especially at the edges of the pictures. This results in the actual covariance matrix being positive semi-definite instead of positive definite. A PD-enforcing scheme could therefore distort the estimation quite heavily.

2. Facial data generating positive semi-definite matrices result in determinants of zero, which there is no inverse, which is needed for the scheme. This can be circumvented by taking Moore-Penrose pseudo-inverses [23], however when the author attempted this work-around, the matrix remained singular to working precision.

3. The scheme requires more processing time.

4. Lastly, in none of the synthetic datasets that parameters were estimated from, did it occur that the diagonal elements had to be changed. No case was noted where the $\hat{C}$-estimator did not generate a positive (semi)-definite covariance matrix.

In this experiment, we attempt to simulate possible dataset errors. The first of these is simply a collection of random outliers, which happen if for example a rogue picture finds its way in the database. The second kind of error that is simulated is an different distribution that is mixed through the data that is to be estimated. This happens if for example a series of pictures is taken with bad lighting.
We ask the following questions that we hope to answer with the two fprms of simulations of database contamination:

1. Will the $\hat{C}$-estimator perform worse under higher covariance? Figure 2 predicts worse results if the covariance is higher. To test this, we will use 100 datapoints and 10 outliers. The outliers are there because the the sample covariance is optimal otherwise and we are interested in robust estimation. We will subsequently higher the covariance from 0 to 0.5 to 0.99.

2. Will the MCD estimator perform worse under influence of more outliers, as predicted by Croux and Haesbroeck [13]. We use the outlier distribution to answer the question, as in reality, one is more likely to encounter a large amount of similarly contaminated samples than encounter a database with many single outliers.

3. Which estimator deals the best with very contaminated samples? For this, we use the single outliers set-up and increase the radius of the circle on which the outliers are placed, thus steadily making the outliers worse and worse.

5.1 Single Outliers Set-Up

The first synthetic test was with **n** randomly generated data following a normal distribution N(0,1) with covariance 0. Then **out** outliers were added to the total data, the outliers were generated as points on a circle with radius R. Then the covariance matrix was determined in three different ways:

1. The sample covariance S
2. The MCD covariance $\hat{\Sigma}_{\text{MCD}}$
3. The $\hat{C}$-estimator covariance $\hat{C}$.

This process was repeated 100 times, every time with new random data and new outliers. The average of the covariance matrix elements are shown below, as are their variances.

Values: n = 100, R = 10 and out = 10.

Table 1: S Average

5.94	0.02
0.02	5.02

Table 2: S Variance

0.02	0.01
0.01	0.02

Table 3: $\hat{\Sigma}_{\text{MCD}}$ Average

1.15	0.03
0.03	1.12

Table 4: $\hat{\Sigma}_{\text{MCD}}$ Variance

0.05	0.02
0.02	0.05

Table 5: $\hat{C}$ Average

1.44	0.03
0.03	1.32

Table 6: $\hat{C}$ Variance

0.05	0.02
0.02	0.06

Values: n = 100, R = 20, out = 10.

Table 7: S Average

21.06	-0.02
-0.02	17.44

Table 8: S Variance

0.02	0.01
0.01	0.01

Table 9: $\hat{\Sigma}_{\text{MCD}}$ Average

1.15	-0.03
-0.03	1.17

Table 10: $\hat{\Sigma}_{\text{MCD}}$ Variance

0.05	0.02
0.02	0.04

Table 11: $\hat{C}$ Average

1.50	-0.04
-0.04	1.34

Table 12: $\hat{C}$ Variance

0.06	0.03
0.03	0.04

Values: n = 100, R = 30 and out = 10.

Table 13: S Average

46.27	-0.004
-0.004	38.05

Table 14: S Variance

0.02	0.01
0.01	0.02

Table 15: $\hat{\Sigma}_{\text{MCD}}$ Average

1.18	0.00
0.00	1.13

Table 16: $\hat{\Sigma}_{\text{MCD}}$ Variance

0.05	0.02
0.02	0.04

Table 17: $\hat{C}$ Average

1.54	0.01
0.01	1.32

Table 18: $\hat{C}$ Variance

0.05	0.03
0.03	0.05

Then the non-diagonal elements of the actual covariance are varied to 0.5 and then to 0.99, since 1 will give a singular matrix. **Values**: n = 100, $\mu = 0$, $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, out = 10.

Table 19: S Average

5.96	0.47
0.47	5.08

Table 20: S Variance

0.02	0.01
0.01	0.02

Table 21: $\hat{\Sigma}_{\text{MCD}}$ Average

1.16	0.59
0.59	1.18

Table 22: $\hat{\Sigma}_{\text{MCD}}$ Variance

0.05	0.03
0.03	0.05

Table 23: $\hat{C}$ Average

1.47	0.73
0.73	1.38

Table 24: $\hat{C}$ Variance

0.05	0.03
0.03	0.06

Values: n = 100, $\mu = 0$, $\Sigma = \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}$, out = 10.

Table 25: S Average

5.95	0.90
0.90	5.04

Table 26: S Variance

0.02	0.02
0.02	0.02

Table 27: $\hat{\Sigma}_{\text{MCD}}$ Average

1.16	1.15
1.15	1.16

Table 28: $\hat{\Sigma}_{\text{MCD}}$ Variance

0.05	0.05
0.05	0.05

Table 29: $\hat{C}$ Average

1.45	1.40
1.40	1.37

Table 30: $\hat{C}$ Variance

0.06	0.06
0.06	0.06

5.2 Outlier Distribution Set-Up

To simulate contamination of the database with another set of data, we will once again vary the amount of outliers, but this time they are distributed following $\text{N}(\mu_{out}, \Sigma_{out})$.

Values: n = 100, μ_{out} = (5,5), $\Sigma_{out} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, out = 10.

Table 31: S Average

3.07	2.11
2.11	3.11

Table 32: S Variance

0.11	0.05
0.05	0.08

Table 33: $\hat{\Sigma}_{\text{MCD}}$ Average

1.14	0.00
0.00	1.14

Table 34: $\hat{\Sigma}_{\text{MCD}}$ Variance

0.04	0.01
0.01	0.04

Table 35: $\hat{C}$ Average

1.47	0.23
0.23	1.44

Table 36: $\hat{C}$ Variance

0.05	0.02
0.02	0.05

Values: n = 100, μ_{out} = (5,5), $\Sigma_{out} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, out = 20.

Table 37: S Average

4.54	3.52
3.52	4.51

Table 38: S Variance

0.17	0.06
0.06	0.18

Table 39: $\hat{\Sigma}_{\text{MCD}}$ Average

1.31	-0.01
-0.01	1.33

Table 40: $\hat{\Sigma}_{\text{MCD}}$ Variance

0.05	0.02
0.02	0.07

Table 41: $\hat{C}$ Average

1.98	0.49
0.49	2.01

Table 42: $\hat{C}$ Variance

0.09	0.03
0.03	0.10

Values: n = 100, μ_{out} = (5,5), $\Sigma_{out} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, out = 40.

Table 43: S Average

6.12	5.13
5.13	6.11

Table 44: S Variance

0.11	0.06
0.06	0.15

Table 45: $\hat{\Sigma}_{\text{MCD}}$ Average

1.74	0.01
0.01	1.73

Table 46: $\hat{\Sigma}_{\text{MCD}}$ Variance

0.08	0.04
0.04	0.08

Table 47: $\hat{C}$ Average

3.07	1.04
1.04	3.06

Table 48: $\hat{C}$ Variance

0.18	0.05
0.05	0.15

6 Real Data Test

Spreeuwers [3] designed a classifier to decide if a probe sample X belongs to the same class c_i as a gallery sample Y. This classifier uses linear discriminant analyses (LDA) [24] as well as principle component analysis (PCA). The classifier needs a *within-class-mean* μ_w, a *total-mean* μ_t and a *within-class* and *total* covariance matrix respectively denoted as C_w and C_t. From these matrices and vectors, a statistic is derived that assigns a value to the samples X and Y. It is also possible to enter more probe samples or more gallery samples or both. If the statistic has a value above 1, the chance that the samples are from the same class is higher than the opposite and vice-versa for a value below 1. If the statistic assigns 1, the chance that the samples are from the same class is just as high as the chance that they are not.

The test: 100 pictures were taken from a database, of which 1 was an obvious outlier. These pictures had 87x75=6525 pixels ranging from 0 to 255 depending on their black-white scale. These 100 pictures were used to determine two covariance matrices to estimate C_t, one normal sample covariance matrix and the other the matrix generated by the robust $\hat{C}$-estimator. Remark: it took 8 hours to process the robust estimator. There exists a faster algorithm for the estimation of Q_n made by Croux and Rosseeuw [25]. The author recommends using this version, it improves the time from O(n^2) to O(n log(n)). Remark 2: The MCD-estimator was not used here as the code required too many samples.
Then a selection of 20 pictures was made, 19 of which belonged to the same person and 1 obvious outlier. These 20 pictures were used to estimate C_w. Similarly to estimating C_t, there were a robust version and a non-robust version of this covariance matrix too. In addition, anywhere where a location estimate was needed, the median was used.
Subsequently, 2 pictures of the same person were fed to the classifier, one as the probe and one as the gallery. The classifier first assigned a score on basis of the non-robust covariance matrices, then assigned a score on basis of the robust covariance matrices.
Result: The score, using non-robust matrices: 0.00459. Using robust matrices: 1015.98.

7 Discussion

The theoretical part of the paper had 4 main results:

1. Conditioning the estimate to become PD;
2. Proving the $\hat{C}$-estimator is B-robust;
3. Evaluating the asymptotic variance of the $\hat{C}$-estimator at Gaussian distributions;
4. Evaluating the Gaussian efficiency of $\hat{C}$-estimator.

The first result does not carry a lot of importance: as discussed in section 5, it is not very useful in the context of facial recognition. However the result is still valid for possible other cases where an estimate has to be PD and the statistician is willing to let go of unbiasedness for the diagonal elements.
The second result is necessary to even consider using the $\hat{C}$-estimator. Without robustness properties, we shall not consider it as a candidate. What this result is still lacking however is a concrete value for the gross-error sensitivity to see exactly how well the estimator resists point mass contaminations. This is still an open question.
The third result is in agreement with the findings of Croux and Rosseeuw. We can conclude that we can build a robust estimator that achieves a $V(\hat{C}, \Phi)$ of 0.6089 for every covariance and not only for variance alone. It is worthwhile to research further applications of this estimator. The small peaks are a bit strange, the author suspects numerical computing errors, but for a flatlined asymptotic variance, the results fully agree with Croux and Rosseeuw.
The fourth result has room for interpretation. For uncorrelated random variables X and Y, the $\hat{C}$-estimator is virtually the same as Q_n as it can estimate their variances with the same high efficiency. However, as soon as X and Y become correlated, the efficiency drops heftily and the estimator becomes fairly weak. This partially has to do with very high values for the Fisher information on covariances that go to 1 or -1. On the topic of the Fisher information, it turned out we had to consider Cov(X,Y) but also Cov(Y,X). It evokes a question if we have to consider all parameters Θ_i that have the same value when evaluating the (joint) Fisher infomation.

The experimental part of the paper asked 3 ques-

tions, which can now be answered qualitatively. The $\hat{C}$-estimator did not perform worse when the covariance was increased. From tables 6, 24 and 30, we can see that the variance only marginally increased. In fact, increasing the covariance did not make any of the three estimators perform worse than they already did.
The MCD estimator indeed showed an increase in variance when the amount of outliers was increased. However, the other two estimators performed far worse and had a much larger increase in variance. It makes sense that all estimators suffer from an increase in contamination, but the relation between the data-outlier ratio has not been considered for the $\hat{C}$-estimator in the theoretical section and could make an interesting follow-up research.
Then from tables 1 to 18, we can clearly see that the MCD-estimator performs best under progressively worse outliers. Even for R = 30, the average covariance matrix did almost not differ from the previous two averages. The sample covariance matrix becomes no longer a valid choice while the $\hat{C}$-estimator does not suffer much but permanently starts too high. Figure 3 and those above it show a visual representation of the ellips described by the different estimators for the covariance.

Lastly a very important conclusion must be drawn: from this paper we can conclude that robust estimation helps in the facial recognition branch of the biometrics. Whilst the author was unable to implement the MCD-estimator for the LDA-classifier test, the implementation of the $\hat{C}$-estimator shows immediately a positive result. The author however is aware that in practical cases, the test is ran multiple times for many pairs of faces, same class or not. He recommends starting here for future testing of robust estimates.

8 Acknowledgement

The author would like to think dr. ir. L.J. Spreeuwers for his supervision and help. He would also like to thank dr. ir. J. Goseling for his role as the external supervisor. Then a special thanks to the reviewing bachelor committee consisting of the two aforemented gentlemen and prof. dr. ir. R.N.J. Veldhuis.

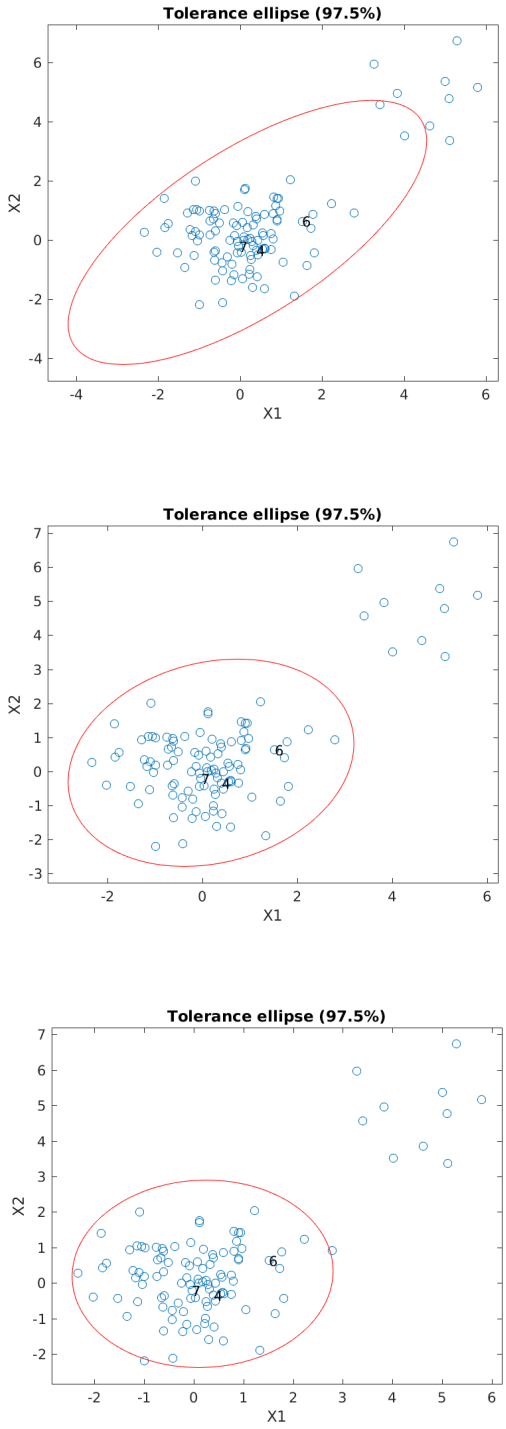

Figure 3: 97.5% confidence ellipses for (top to bot) S, $\hat{C}$, MCD

References

[1] I. van der Ploeg, "Genetics, biometrics and the informatization of the body," *ANNALI-ISTITUTO SUPERIORE DI SANITA*, vol. 43, no. 1, p. 44, 2007.

[2] L. Spreeuwers, "Fast and accurate 3d face recognition," *International journal of computer vision*, vol. 93, no. 3, pp. 389–414, 2011.

[3] L. Spreeuwers, "Breaking the 99% barrier: optimisation of three-dimensional face recognition," *IET biometrics*, vol. 4, no. 3, pp. 169–178, 2015.

[4] D. Peña and F. J. Prieto, "Multivariate outlier detection and robust covariance matrix estimation," *Technometrics*, vol. 43, no. 3, pp. 286–310, 2001.

[5] F. R. Hampel, "Contribution to the theory of robust estimation," *Ph. D. Thesis, University of California, Berkeley*, 1968.

[6] R. Gnanadesikan and J. R. Kettenring, "Robust estimates, residuals, and outlier detection with multiresponse data," *Biometrics*, pp. 81–124, 1972.

[7] R. A. Maronna and V. J. Yohai, "Robust estimation of multivariate location and scatter," *Wiley StatsRef: Statistics Reference Online*, 1976.

[8] M. Hubert and M. Debruyne, "Minimum covariance determinant," *Wiley interdisciplinary reviews: Computational statistics*, vol. 2, no. 1, pp. 36–43, 2010.

[9] D. Donoho, I. Johnstone, P. Rousseeuw, and W. Stahel, "Discussion: projection pursuit," *The Annals of Statistics*, vol. 13, no. 2, pp. 496–500, 1985.

[10] S. Van Aelst and P. Rousseeuw, "Minimum volume ellipsoid," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 1, pp. 71–82, 2009.

[11] F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel, "Robust statistics, j," *Wiley& Sons, New York*, 1986.

[12] G. J. McLachlan, "Mahalanobis distance," *Resonance*, vol. 4, no. 6, pp. 20–26, 1999.

[13] C. Croux and G. Haesbroeck, "Influence function and efficiency of the minimum covariance determinant scatter matrix estimator," *Journal of Multivariate Analysis*, vol. 71, no. 2, pp. 161–190, 1999.

[14] R. Couillet, F. Pascal, and J. W. Silverstein, "Robust estimates of covariance matrices in the large dimensional regime," *IEEE Transactions on Information Theory*, vol. 60, no. 11, pp. 7269–7278, 2014.

[15] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[16] P. J. Huber, *Robust statistical procedures*. SIAM, 1996.

[17] G. T. Gilbert, "Positive definite matrices and sylvester's criterion," *The American Mathematical Monthly*, vol. 98, no. 1, pp. 44–46, 1991.

[18] J. J. Sylvester, "Xxxvii. on the relation between the minor determinants of linearly equivalent quadratic functions," *Philosophical Magazine Series 4*, vol. 1, no. 4, pp. 295–305, 1851.

[19] P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *Journal of the American Statistical association*, vol. 88, no. 424, pp. 1273–1283, 1993.

[20] A. M. Pires and J. A. Branco, "Partial influence functions," *Journal of Multivariate Analysis*, vol. 83, no. 2, pp. 451–468, 2002.

[21] K. Long, "Gateaux differentials and frechet derivatives." http://www.math.ttu.edu/ klong/5311-spr09/diff.pdf.

[22] P. Zegers, "Fisher information properties," *Entropy*, vol. 17, no. 7, pp. 4918–4939, 2015.

[23] J. S. Golan, "Moore–penrose pseudoinverses," *The Linear Algebra a Beginning Graduate Student Ought to Know*, pp. 441–452, 2012.

[24] B. D. Ripley, *Modern applied statistics with S*. Springer, 2002.

[25] C. Croux and P. J. Rousseeuw, *Time-efficient algorithms for two highly robust estimators of scale*. Universitaire Instelling Antwerpen. Department of Mathematics, 1992.